Manual de usuario para Jwire 1.0

Luis Alberto García Hernández

Información de los Autores:

Luis Alberto García Hernández
Ingeniero en Informática
luisalbertogh@hotmail.com
http://it.luisalbertogh.net

Este documento puede ser libremente distribuido.
(c) 2010 Luis Alberto García Hernández

Resumen

WIRE es un *crawler* o araña web desarrollado por un equipo de la Universidad de Chile bajo licencia GPL, es decir, de uso público. El crawler es capaz por si sólo de rastrear y extraer gran cantidad de información de Internet o de una Intranet, accediendo a múltiples sitios web. Digamos que es capaz de rastrear un dominio de un país completo si se le dan los recursos computacionales y el tiempo necesario.

Este crawler es utilizado por profesores e investigadores de la Universidad de Salamanca en la realización de sus trabajos dentro de diferentes campos de la Informática, como pueden ser la Minería Web y la Cibermetría. Debido a la complejidad de las operaciones realizadas por el crawler y a la ausencia de interfaces gráficas, el uso de WIRE puede llegar a ser un tanto engorroso y complicado para usuarios poco expertos. Además, en el caso de los crawlers en general, y de WIRE en particular, la cantidad de información extraida de las redes rastreadas es, en la mayoría de los casos, muy importante. Analizar dicha información es un trabajo complejo sin las técnicas y las herramientas adecuadas.

El objetivo de este proyecto es analizar las diferentes técnicas de visualización que se pueden aplicar al análisis de los datos obtenidos por un crawler, en este caso el crawler WIRE, y desarrollar una serie de herramientas de visualización que implementen dichas técnicas con el objetivo de verificar y validar su utilidad. Este conjunto de herramientas se desarrollará en un entorno gráfico de uso para WIRE, denominado **Jwire**. De ahí que el título original del proyecto haya pasado de ser "**WIRE, un crawler multifunción**", al actual. Dotando a WIRE de una interfaz gráfica desde la cual se puedan hacer uso de dichas herramientas de visualización, junto a las funcionalidades propias de WIRE, pero de una forma sencilla y rápida, haciéndolo así más asequible para cualquier tipo de usuario, automatizando tareas y en definitiva añadiendo un valor adicional a la aplicación ya existente.

A continuación se presenta una breve descripción gráfica de las funcionalidades de Jwire en su versión 1.0. Para más información acerca del funcionamiento de Jwire, consultar el manual técnico.

Índice

Índice de figuras III

1. Requisitos 1

2. Instalación y ejecución 1

3. Módulo de configuración 2
3.1. Editar wire.conf 2
3.2. Editar fichero de urls 3

4. Módulo de ejecución 5
4.1. Ejecución de comandos WIRE 5
4.1.1. Crear varias listas de ejecución 6
4.1.2. Visualizar y exportar la salida estandar de los comandos . . . 6
4.1.3. Barra de controles 7
4.2. Gestión de informes WIRE 8

5. Módulo de visualización 9
5.1. Grafos dirigidos por fuerzas 11
5.2. Grafos multicapa 13
5.3. Herramientas para grafos 16

Índice de figuras

1. Edición de wire.conf . . . 2
2. Barra de comandos . . . 3
3. Propiedades de wire.conf . . . 3
4. Edición de fichero de urls . . . 4
5. Barra de comandos . . . 4
6. Tabla de urls . . . 5
7. Módulo de ejecución . . . 5
8. Ejecución de comandos . . . 6
9. Varias listas a la vez . . . 7
10. Salida y error estandar . . . 7
11. Barra de controles . . . 7
12. Configuración de aplicaciones . . . 8
13. Acceso a ficheros . . . 8
14. Módulo de visualización . . . 10
15. Barra de herramientas de visualización . . . 10
16. Visaulizar datos en Jwire . . . 11
17. Grafo dirigido por fuerzas . . . 12
18. Información sobre el grafo . . . 12
19. Agrupación jerárquica del grafo multicapa . . . 13
20. Nivel 0 . . . 14
21. Nivel 1 . . . 14
22. Nivel 2 . . . 15
23. Barra de herramientas para grafos . . . 16
24. Grafo radial . . . 17
25. Treemap . . . 18
26. Panel de fuerzas y buscador de nodos . . . 18
27. Datos sobre los nodos . . . 19

1. Requisitos

Jwire está implementado en **Java** estandar con el SDK de Sun, versión 6 update 18. Jwire requiere por lo tanto del JRE de Java, que ha de estar instalado previamente en el sistema sobre el que se vaya a ejecutar, en su versión 5 o superior. Se recomienda la versión 6 update 10 como mínimo para que Jwire explote sus características al máximo.

2. Instalación y ejecución

Descomprimir el archivo Jwire (`zip` o `tar`) en el directorio desde el cual se vaya a ejecutar. El archivo de Jwire contiene todos los ficheros y directorios que Jwire necesita. Jwire considerará el directorio desde el que se ejecuté como su directorio raíz en el cual almacenar la información necesaria para su correcta ejecución.

Para ejecutar Jwire, lanzar el script de ejecución correspondiente al sistema operativo en el que se instaló (`.sh` para Unix/Linux o `.bat` para Windows)

3. Módulo de configuración

3.1. Editar wire.conf

1. Seleccionar la pestaña de **Configuración**
2. Hacer click sobre el botón con nombre **wire.conf**

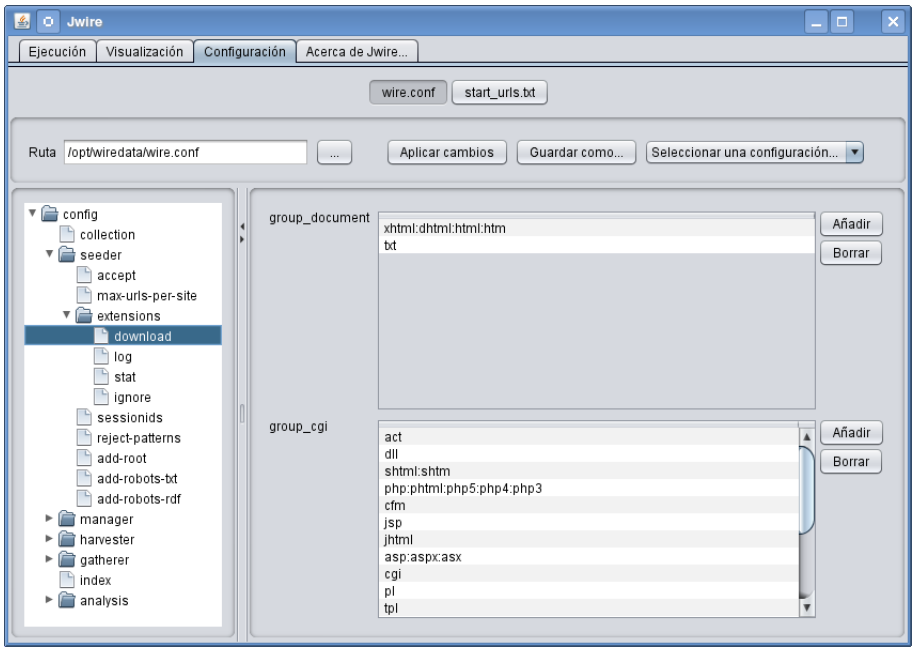

Figura 1: Edición de wire.conf

- Usar el campo ruta para seleccionar un nuevo fichero de configuración
- Usar **Aplicar cambios** para guardar los cambios realizados
- Usar **Guardar como...** para almacenar los cambios a un fichero sin aplicarlos a wire.conf
- Usar el combo **Seleccionar una configuración...** para cargar los datos almacenados en fichero para su posterior modificación en Jwire

Figura 2: Barra de comandos

- Seleccionar los nodos del árbol que representan cada uno de los campos a editar en wire.conf
- En la parte derecha aparecen los valores que se pueden editar para cada campo seleccionado, Existen diferentes tipos de campos. Cada tipo de campo posee diferentes controles, algunos son valores numéricos, otros son tablas, otros son campos de texto, etc.

Figura 3: Propiedades de wire.conf

3.2. Editar fichero de urls

1. Seleccionar la pestaña de **Configuración**
2. Hacer click sobre el botón con nombre `start_urls.txt`

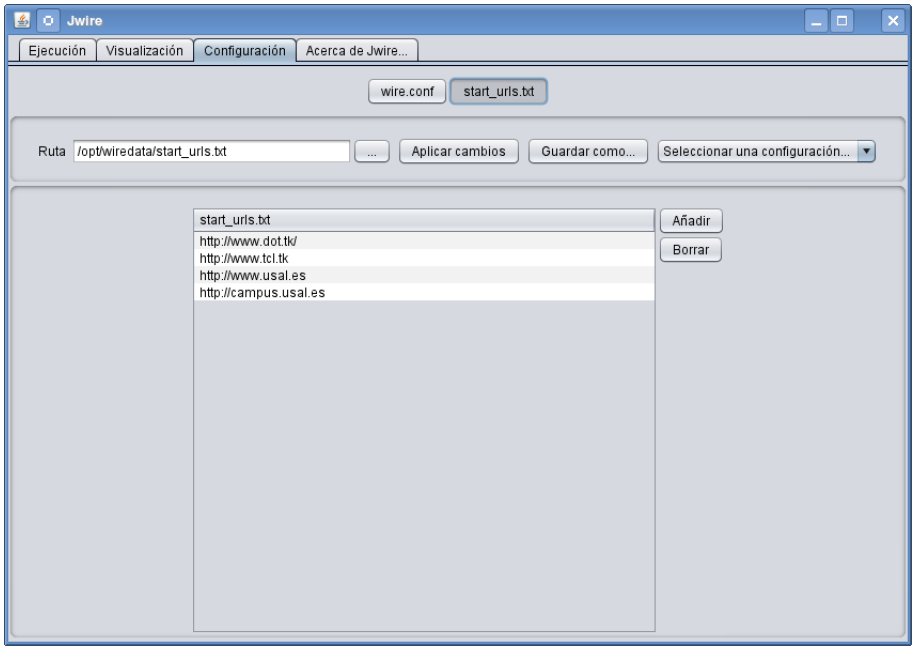

Figura 4: Edición de fichero de urls

- Usar el campo ruta para seleccionar un nuevo fichero de urls
- Usar **Aplicar cambios** para guardar los cambios realizados
- Usar **Guardar como...** para almacenar los cambios a un fichero sin aplicarlos a wire.conf
- Usar el combo **Seleccionar una configuración...** para cargar los datos almacenados en fichero para su posterior modificación en Jwire

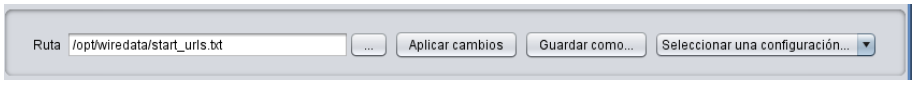

Figura 5: Barra de comandos

- Usar la tabla para añadir, eliminar y editar **urls** para su uso con Jwire

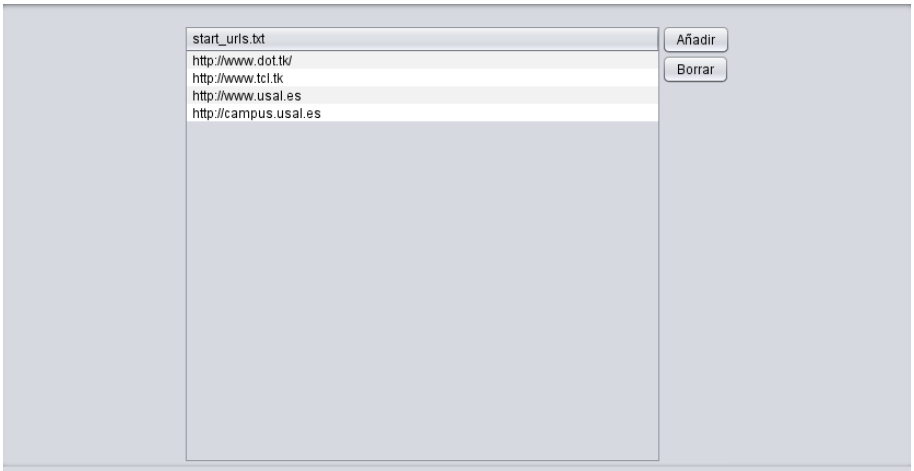

Figura 6: Tabla de urls

4. Módulo de ejecución

4.1. Ejecución de comandos WIRE

Tras seleccionar la pestaña de **Ejecución**...

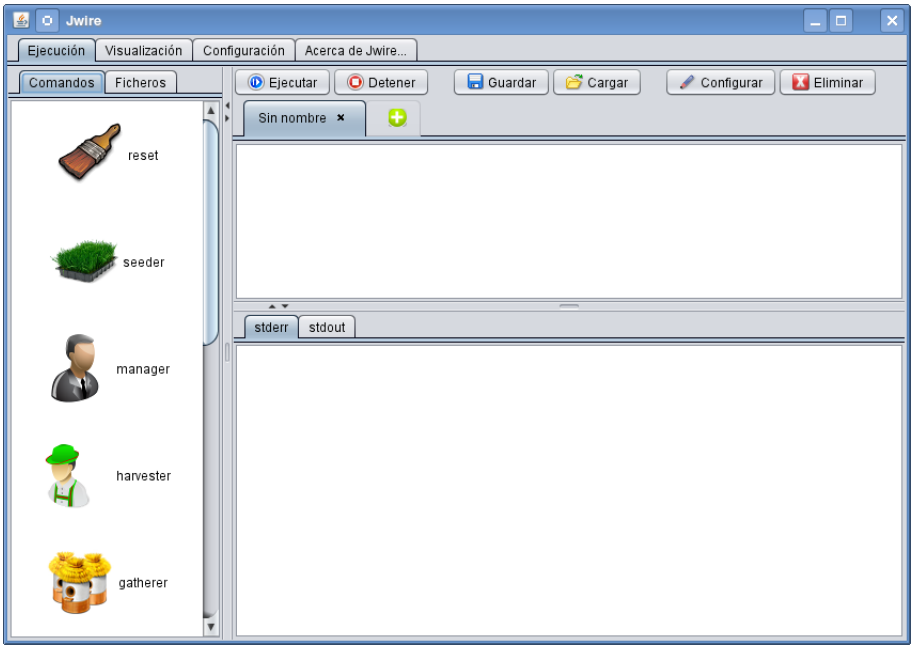

Figura 7: Módulo de ejecución

1. Seleccionar el comando WIRE de la lista de comandos
2. Arrastrar los comandos seleccionados a la lista de ejecución
3. Opcionalmente, configurar cada uno de los parámetros de ejecución de cada uno de los comandos, usando el botón **Configurar** de la barra de control

4. Ejecutar la lista seleccionada. Los comandos de la lista se ejecutarán por orden de izquierda a derecha. El comando en ejecución aparece con el fondo en rojo mientras se está ejecutando

5. En la parte inferior de la ventana aparecen la salida y el error standar de la ejecución de los comandos

Para más información acerca de la barra de control del módulo de ejecución, dirigirse a la sección 4.1.3.

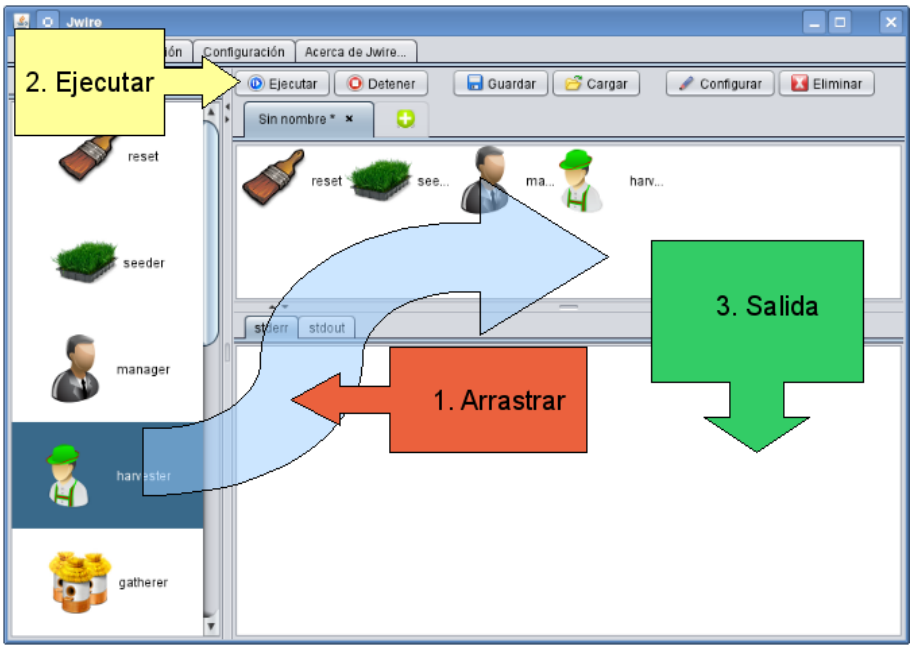

Figura 8: Ejecución de comandos

4.1.1. Crear varias listas de ejecución

Se pueden manejar varias listas de ejecución al mismo tiempo, creando nuevas pestañas en la parte superior del recuadro central. Basta con clicar sobre el círculo verde con la cruz blanca que aparece junto a las pestañas existentes. Cada vez que se crea una nueva pestaña, se le asigna el título **Sin nombre** por defecto. Este se puede cambiar al guardar las listas de ejecución, como veremos a continuación.

4.1.2. Visualizar y exportar la salida estandar de los comandos

Existen dos paneles diferentes asociados a cada lista de ejecución, que muestran respectivamente la salida y el error por consola de cada uno de los comandos ejecutados en la lista, El contenido de estos paneles se puede exportar a un fichero de texto haciendo **click con el botón derecho del ratón** para mostrar un **menu pop-up**, y después escogiendo la opción **Exportar** e indicando el **fichero de destino**.

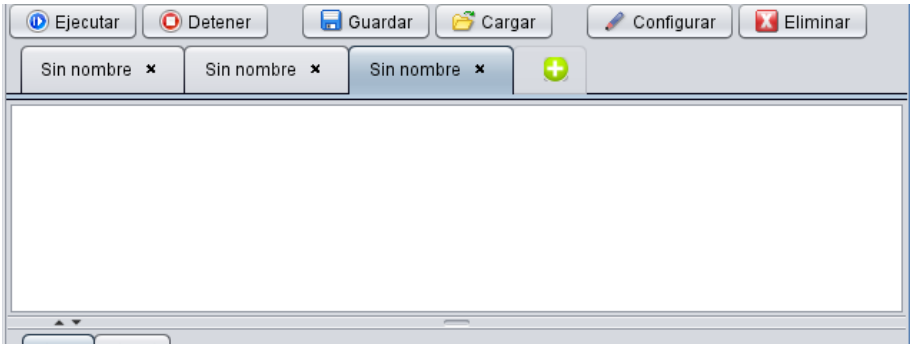

Figura 9: Varias listas a la vez

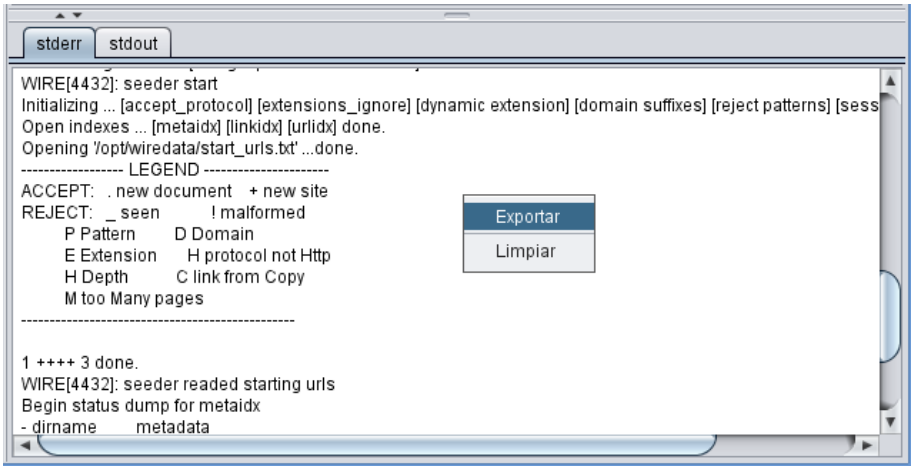
stderr
stdout
WIRE[4432]: seeder start
Initializing ... [accept_protocol] [extensions_ignore] [dynamic extension] [domain suffixes] [reject patterns] [sess
Open indexes ... [metaidx] [linkidx] [urlidx] done.
Opening '/opt/wiredata/start_urls.txt' ...done.
---------- LEGEND ----------
ACCEPT: . new document + new site
REJECT: _ seen ! malformed
P Pattern D Domain
E Extension H protocol not Http
H Depth C link from Copy
M too Many pages
1 ++++ 3 done.
WIRE[4432]: seeder readed starting urls
Begin status dump for metaidx
- dirname metadata
Exportar
Limpiar

Figura 10: Salida y error estandar

4.1.3. Barra de controles

Existen diferentes controles dentro de la barra situada en la parte superior del panel central de ejecución de Jwire. Dichos controles tienen las siguientes funciones:

Figura 11: Barra de controles

- **Ejecutar**. Ejecución de la lista de comandos
- **Detener**. Detiene la ejecución de la lista de comandos, si es posible
- **Guardar**. Guarda la lista de comandos selecciona a fichero, para su posterior reutilización dentro de Jwire
- **Cargar**. Carga en Jwire una lista de comandos desde fichero, y que ha sido previamente guardada desde Jwire
- **Configurar**. Permite configurar los parámetros de ejecución de cada comando WIRE, según las necesidades de cada uno

- **Eliminar**. Elimina de la lista de ejecución el comando seleccionado

4.2. Gestión de informes WIRE

Desde Jwire se pueden abrir los informes generados por los comandos WIRE. Jwire hace uso de las aplicaciones existentes en el sistema operativo sobre el que se instaló para abrir dichos ficheros. Por lo tanto, antes de hacer uso de esta funcionalidad, es necesario configurar Jwire para indicarle con que aplicaciones se abriran los distintos tipos de ficheros.

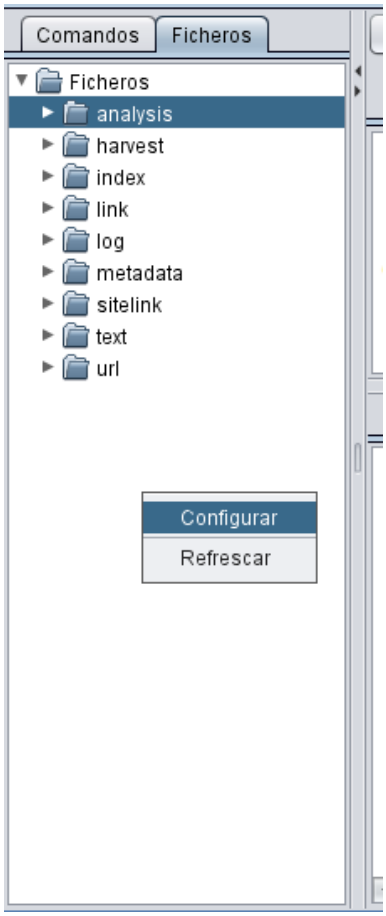

Figura 12: Configuración de aplicaciones

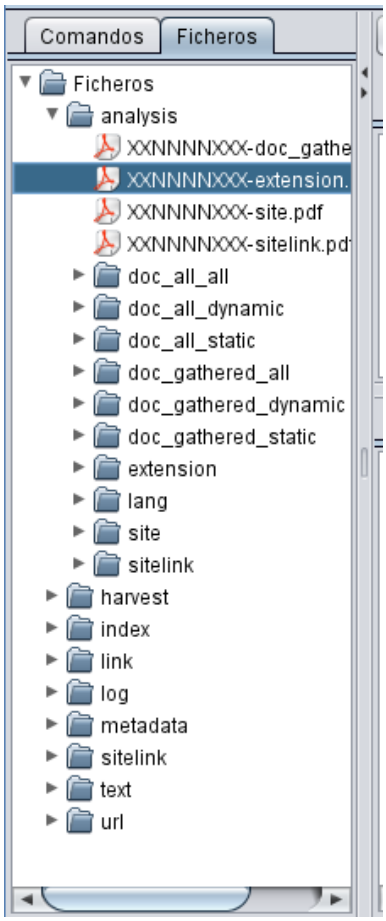

Figura 13: Acceso a ficheros

Para ello:

1. Desde la pestaña de ejecución, se selecciona la pestaña **Ficheros** del panel izquierdo de la ventana.

2. Se muestra un árbol jerárquico con la estructura de directorios y ficheros creada por WIRE para almacenar los informes generados por sus comandos. Haciendo click con el botón derecho del ratón sobre el panel, se muestra un menu pop-up. Escogiendo la opción **Configurar** se abre un cuadro de diálogo que contiene

una tabla. Dicha tabla almacena en dos columnas las rutas de las aplicaciones a usar y las extensiones de los ficheros asociadas a dichas aplicaciones. Usando los botones del cuadro de diálogo o hacindo doble click sobre las filas de la tabla, se pueden editar, añadir o eliminar nuevas asociaciones. Si se indica el valor ***** como extensión, la aplicación asociada se usará para abrir cualquier fichero que no tenga otra asociación establecida en la tabla

Una vez configurado Jwire, se pueden abrir los ficheros generados por WIRE de la siguiente manera:

1. Desde la pestaña de ejecución, se selecciona la pestaña **Ficheros** del panel izquierdo de la ventana.
2. Se muestra un árbol jerárquico con la estructura de directorios y ficheros creada por WIRE para almacenar los informes generados por sus comandos. Navegar dentro de esta estructura hasta encontrar el fichero buscado y hacer doble-click sobre él para abrirlo en una aplicación aparte

Jwire busca los ficheros generados por WIRE en la localización escogida por éste para almacenarlos, y que parte del directorio escogido por WIRE para almacenar los datos. Para configurar la ruta de dicho directorio, es necesario modificar el fichero **wire.conf** y la **variable de entorno** de WIRE.

Por último, se puede refrescar el contenido del árbol de directorios de informes de Jwire, haciendo uso de la opción **Refrescar** del menu pop-up de la pestaña de **Ficheros**.

5. Módulo de visualización

Las herramientas de visualización y análisis visual de Jwire se encuentran disponibles activando la pestaña **Visualización**.

Jwire aporta a WIRE un conjunto de herramientas de visualización que implementan la mayoría de las técnicas estandar para visualizar y analizar datos extraidos de una red, en particular con el uso de **grafos web** y **treemaps**. El módulo de visualización hace uso de los comandos WIRE usados para formatear y generar los datos extraidos por WIRE, y posteriormente usa las técnicas de visualización recomendadas para mostrar esos datos por pantalla en forma de grafos web y otras visualizaciones útiles a la hora de analizar los datos obtenidos.

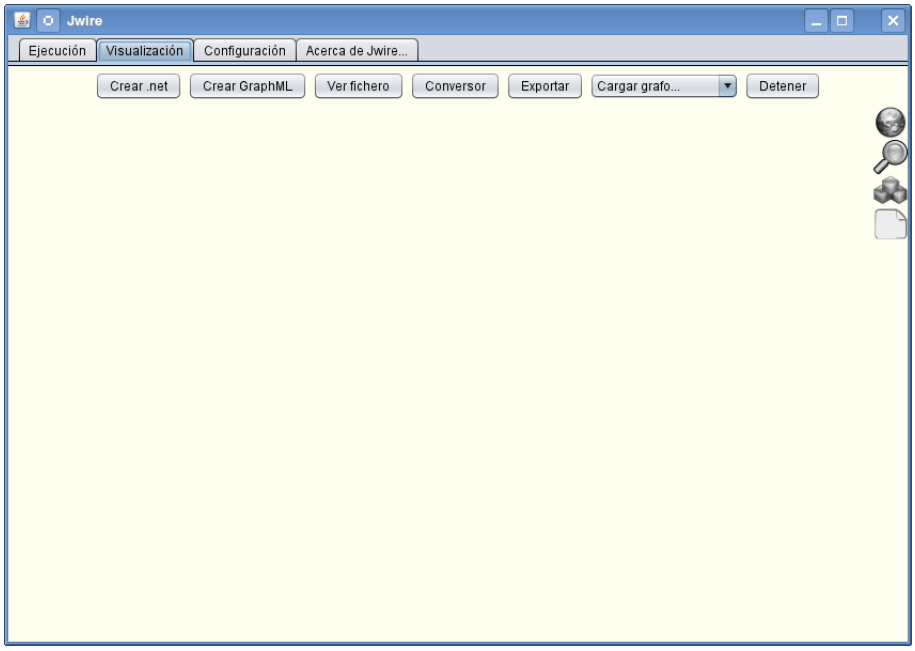

Figura 14: Módulo de visualización

Existe una barra de herramientas en forma de una secuencia de botones, que ofrecen diferentes funcinalidades. Éstas son las siguientes:

Figura 15: Barra de herramientas de visualización

- **Crear .net**. Genera un fichero con formato .net (Pajek) con los datos extraidos en la última ejecución de WIRE. Esto es útil si se quieres visualizar o manipular dichos datos con aplicaciones como Pajek u otras que sólo funcionan con el formato .net
- **Crear GraphML**. Genera un fichero XML con el esquema de GraphML para grafos web. Este es un paso indispensable si se quieren visualizar en Jwire los datos generados por WIRE
- **Ver ficheros**. Abre en una ventana aparte el fichero seleccionado, y lo muestra en modo texto
- **Conversor**. Convierte ficheros de formato .net a GraphML y viceversa
- **Exportar**. Exporta el grafo generado en el área de visualización a una imagen
- **Cargar grafo...**. Una vez que el fichero GraphML ha sido generado, se puede abrir dicho fichero como grafo web, seleccionandolo. Existen dos modos de visualización, el **grafo dirigido por fuerzas**, que muestra la totalidad de los datos generados en un grafo web dirigido por fuerzas, o el **grafo multicapa**,

que muestra los datos distribuidos en una estructura jerárquica que compone un grafo de 3 niveles distintos

- **Detener**. Detiene o reinicia el movimiento del grafo web

La forma de visualizar los datos extraidos por WIRE es la siguiente:

1. Generar el fichero XML con el formato GraphML, haciendo click sobre el botón correspondiente de la barra de herramientas. El fichero ha de tener la extensión `.xml`
2. Cargar el fichero con el grafo. Usar el combo de la barra de herramientas, seleccionando primero el tipo de grafo que se quiere visualizar, y posteriormente el fichero XML correspondiente

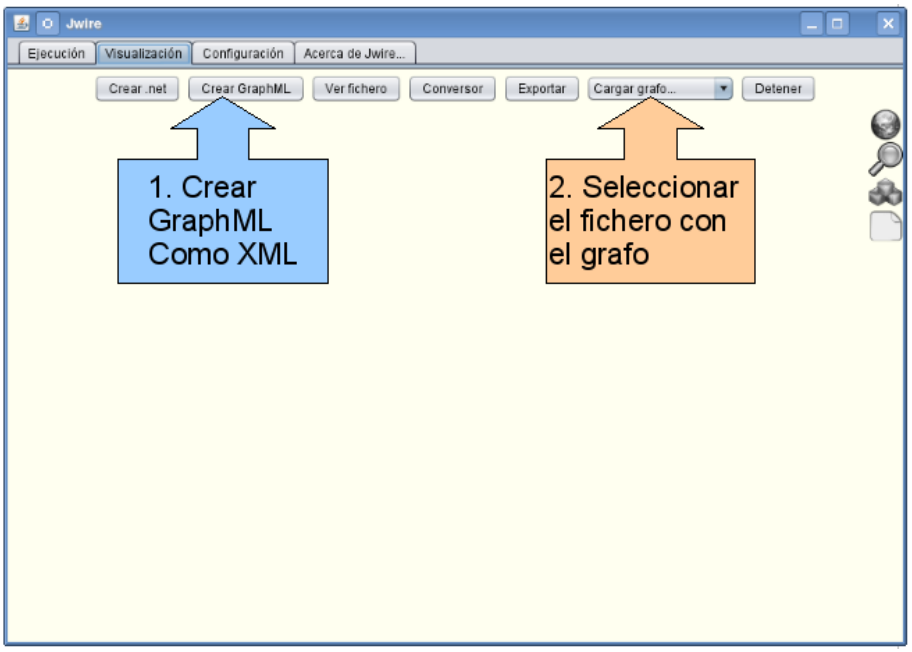

Figura 16: Visaulizar datos en Jwire

5.1. Grafos dirigidos por fuerzas

A continuación se muestra un ejemplo de la visualización de los datos generados por WIRE mediante un grafo dirigido por fuerzas.

Por defecto se muestran las agrupaciones dentro del grafo, mediante las manchas de colores que agrupan los nodos que pertenecen al mismo sitio web. Es decir, cada mancha de color es un único sitio web, del tipo **www.algo.com**.

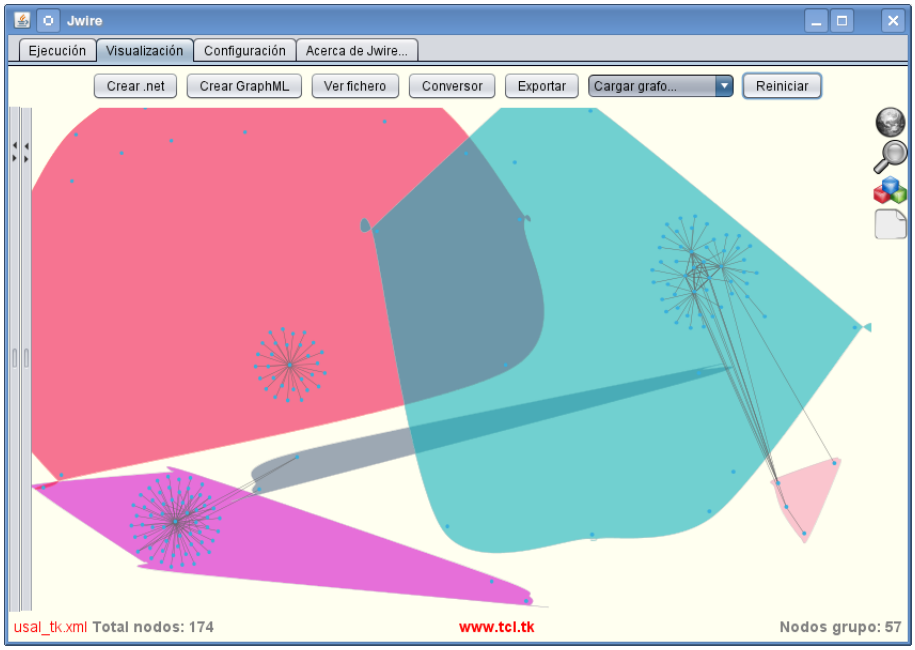

Figura 17: Grafo dirigido por fuerzas

En la parte inferior del área de visualización se muestra información acerca del grafo, y de los nodos o enlaces que se seleccionan posando el puntero del ratón sobre ellos. Normalmente se muestra el nombre del fichero abierto, el número total de nodos, el nombre del nodo o agrupación seleccionado y el número de nodos de la agrupación.

Figura 18: Información sobre el grafo

5.2. Grafos multicapa

El grafo multicapa es un conjunto de grafos web que representan diferentes niveles dentro de una estructura jerárquica de contención. Los datos generados por WIRE se distribuyen en un único nivel, al cual se accede desde otros niveles superiores que se crean de forma artificial y que representan las agrupaciones de esos datos de forma distribuida. El siguiente diagrama representa el mecanismo de contrucción de los niveles del grafo multicapa.

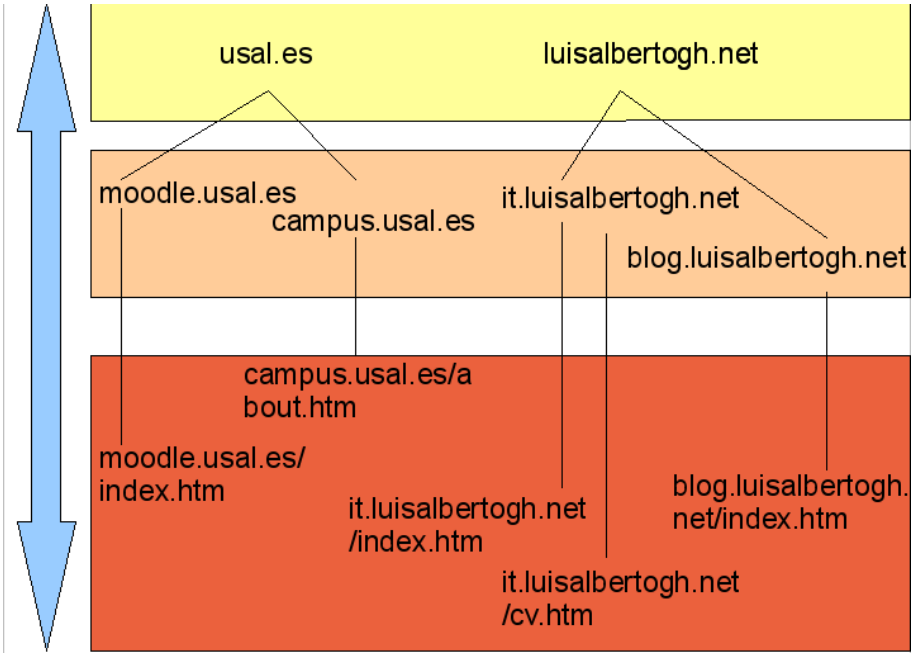

Figura 19: Agrupación jerárquica del grafo multicapa

Los nodos del nivel inferior son los extraidos por WIRE. Jwire muestra por defecto el nivel superior, y se puede acceder a los niveles inferiores haciendo doble click sobre el nodo que representa la agrupación que se quiere explorar.

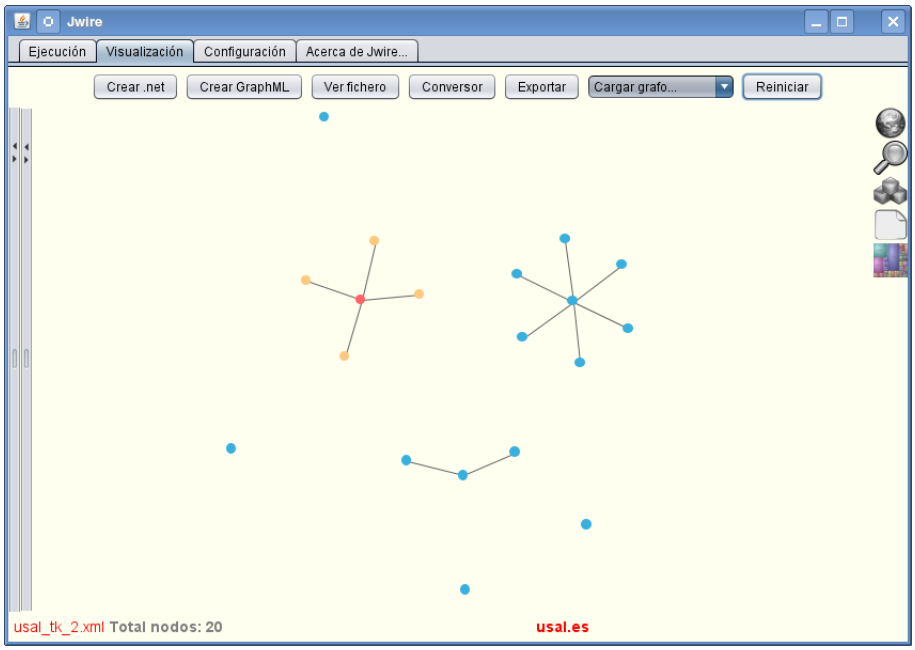

Figura 20: Nivel 0

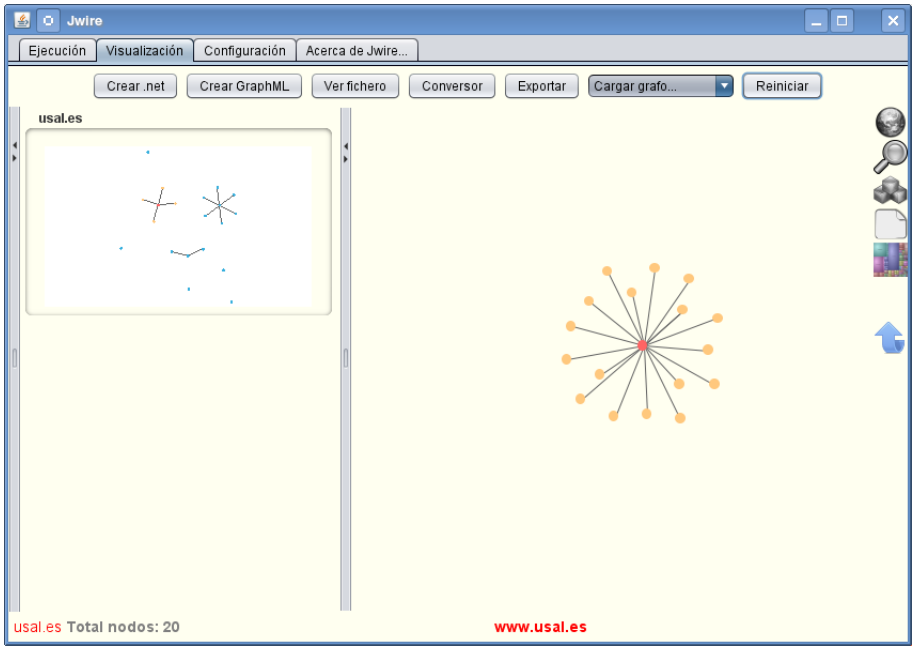

Figura 21: Nivel 1

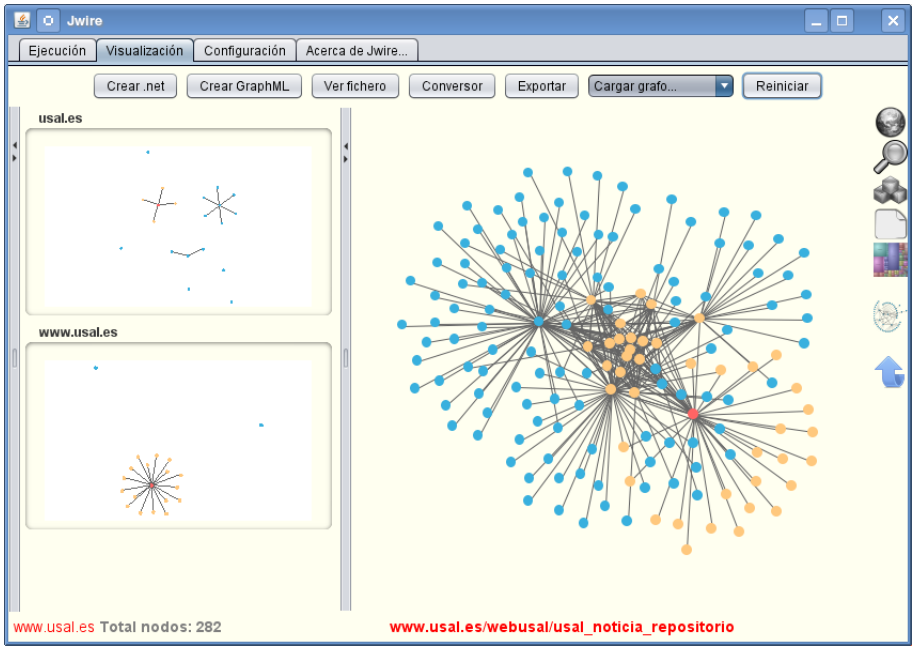

Figura 22: Nivel 2

Como se observa en los niveles 1 y 2, se muestra siempre el nivel anterior en una vista reducida en la parte izquierda, de tal manera que siempre se mantiene visible desde donde se viene, y haciendo click sobre esta vista, se accede de nuevo al nivel correspondiente.

5.3. Herramientas para grafos

Tanto para los grafos dirigidos por fuerzas como para los grafos multicapa, existen una serie de herrmientas implementadas que permiten la manipulación y visualización de los diferentes componentes de los grafos. Estas herramientas implementan la mayoría de las técnicas estandar de visualización de grafos, y son una recomendación a la hora de visualizar los grafos de la mejor forma posible.

La mayoría de estas herramientas se encuentran accesibles desde la barra de tareas que aparece en la parte derecha del área de visualización:

Figura 23: Barra de herramientas para grafos

Esta barra de herramientas no siempre es igual. Sólo aparecen disponibles las opciones existentes para cada tipo de grafo. Las que se muestran en la imagen son las que existen en total, y representan las siguientes funciones, de arriba a abajo:

1. **Vista global**. Muestra sobre el grafo un área reducida que contiene la vista global del grafo, y un área rectangular con la que se puede fijar una zona determinada del grafo que se muestra ampliada en la zona central de la ventana

2. **Lupa**. Usando el cursor del ratón, se puede ampliar una zona determinda del grafo, mientras que el resto se ve más reducido, tal y como se vería usando una lupa

3. **Agrupaciones**. Esta opción sirve para mostrar u ocultar las agrupaciones. En el caso del grafo dirigido por fuerzas, está activada por defecto, y muestra los sitios web como manchas de colores. En el caso del grafo multicapa, está desactivada por defecto, pero se puede activar seleccionando el tipo de agrupación a mostrar, bien sobre dominios de segundo nivel o sobre dominios de primer nivel. En este caso, las agrupaciones se muestran coloreando con el mismo color los nodos que pertenecen al mismo grupo

4. **Tooltips**. Muestra el nombre de los nodos si el puntero del ratón se posa sobre los nodos. Por defecto desactivada

5. **Treemap**. Sólo para grafo multicapa. Muestra en una ventana aparte un treemap del grafo que se muestra en ese momento

6. **Grafo radial**. Sólo para la última capa del grfo multicapa. Muestra en una ventana aparte un grafo radial del grafo abierto en ese momento

7. **Subir nivel**. Sólo para grafos multicapa, sirve para subir de un nivel a otro

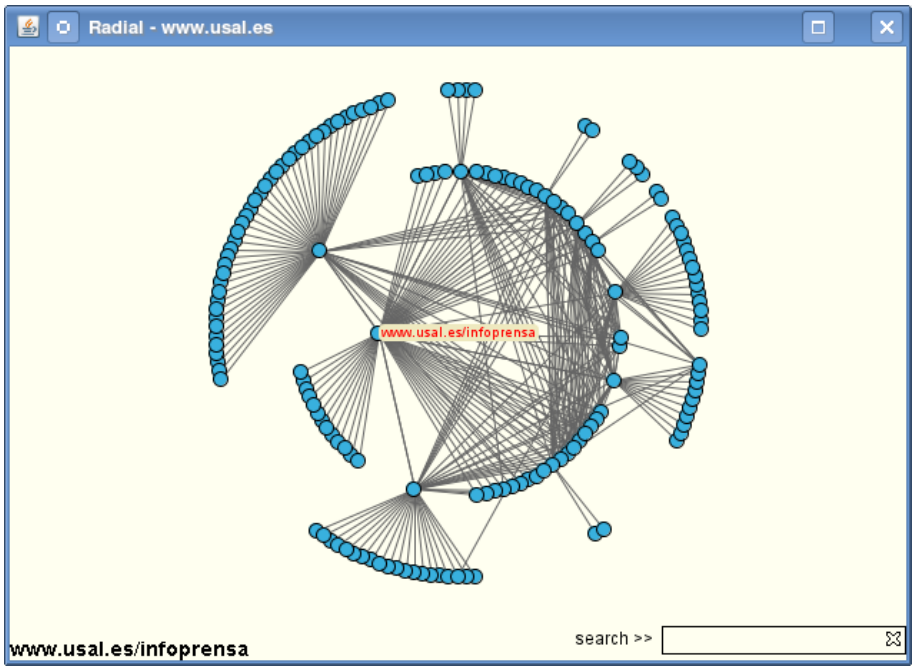

Figura 24: Grafo radial

Existe un panel que permite modificar las propiedades del sistema de fuerzas que controla la disposición de los grafos, así como un buscador de nodos por nombre. El panel se encuentra oculto por defecto, pero se puede mostrar haciendo click o arrastrando sobre las flechas negras, en la parte izquierda de la ventana.

Por último, se han añadido algunas características a los grafo web para facilitar su análisis visual. Los nodos seleccionados con el puntero del ratón se colorean en rojo de forma automática, y en amarillo los nodos enlazados a él directamente. Si se hace click sobre un nodo mientras se pulsa la tecla `SHIFT`, se muestra una pequeña ventana con datos acerca de la conectividad del nodo dentro del grafo (grados, vecinos, etc.). Si se pulsa el nodo mientras se aprieta la tecla `Control`, se abrirá el navegador web por defecto del sistema y se intentará acceder a la dirección web que representa el nodo.

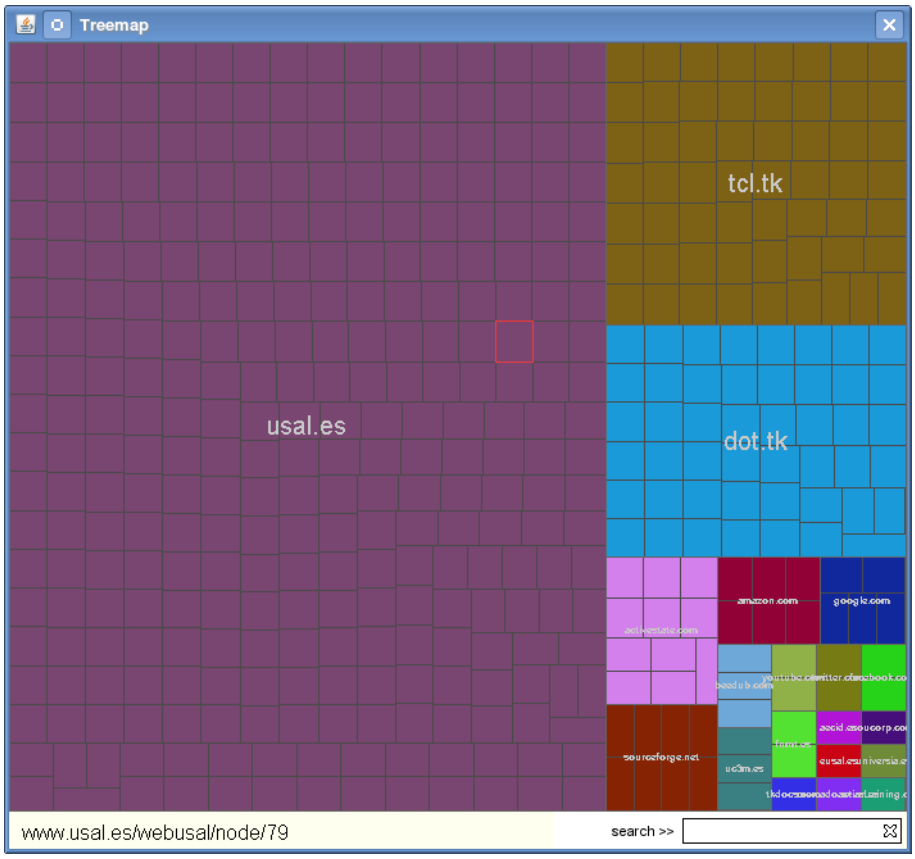

Figura 25: Treemap

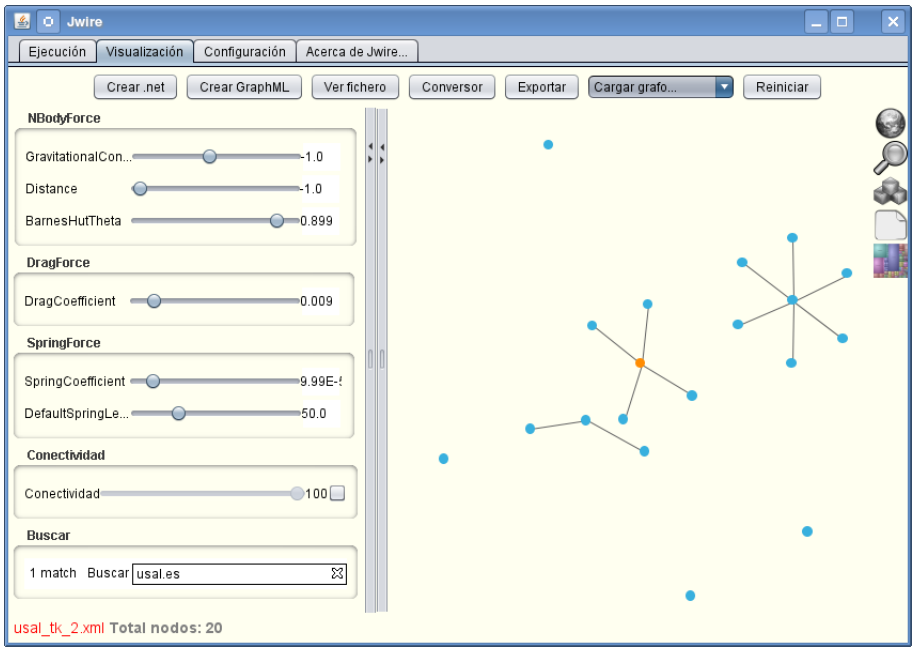

Figura 26: Panel de fuerzas y buscador de nodos

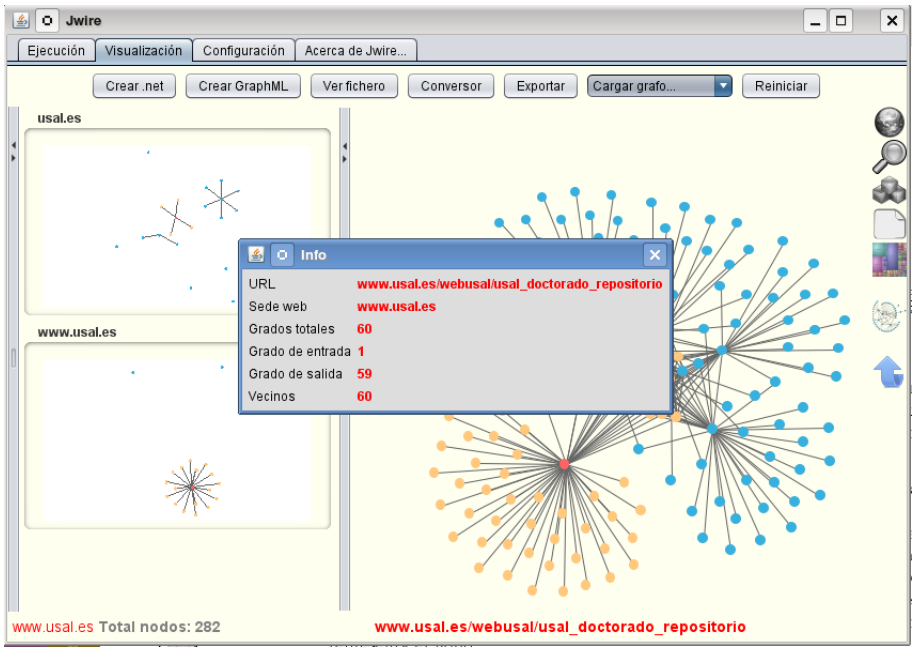

Figura 27: Datos sobre los nodos